

Extending Context Window of Large Language Models via Positional Interpolation

Year: 2023 | Citations: 652 | Authors: Shouyuan Chen, Sherman Wong, Liangjian Chen

Abstract

We present Position Interpolation (PI) that extends the context window sizes of RoPE-based pretrained LLMs such as LLaMA models to up to 32768 with minimal fine-tuning (within 1000 steps), while demonstrating strong empirical results on various tasks that require long context, including passkey retrieval, language modeling, and long document summarization from LLaMA 7B to 65B. Meanwhile, the extended model by Position Interpolation preserve quality relatively well on tasks within its original context window. To achieve this goal, Position Interpolation linearly down-scales the input position indices to match the original context window size, rather than extrapolating beyond the trained context length which may lead to catastrophically high attention scores that completely ruin the self-attention mechanism. Our theoretical study shows that the upper bound of interpolation is at least ~ 600 times smaller than that of extrapolation, further demonstrating its stability. Models extended via Position Interpolation retain its original architecture and can reuse most pre-existing optimization and infrastructure.